AgriGender *Journal of Gender, Agriculture and Food Security*

*AgriGender* 2021; **Volume 06, Issue 02**, 19–40, doi:10.19268/JGAFS.622021.2

# Gender capacity development in agriculture: insights from the GREAT monitoring, learning, and evaluation system

**Cassidy Travis[†*], Elisabeth Garner[◇], Yvonne Pinto[†*], & Godfrey Kayobyo[§]**

[†] *Aline Impact Limited, London, UK.*
[◇] *Department of Global Development, Cornell University, Ithaca, NY, USA.*
[§] *Nkoola Institutional Development Associates (NIDA), Kampala, Uganda.*

[*]Corresponding authors. Email: casstravis@gmail.com, yvonne@alineimpact.com

Capacity development interventions are considered critical entry points for advancing gender equality in agricultural research systems. However, the impacts of capacity development programs are often difficult to track. Academic reviews highlight that insufficient attention is paid to the suitability of gender training programs to increase capacity and limited evidence is available on their longer-term impacts. This article proposes a systematic approach to monitoring, learning, and evaluation (MLE) of gender training programs, which was developed over a five-year period to assess the Gender-Responsive Researchers Equipped for Agricultural Transformation (GREAT) training program. Findings reveal the importance of not only tracking how trainings build technical knowledge but also capturing how trainees are empowered or limited in their efforts to apply gender-responsive practices in diverse environments. The article demonstrates the value of capturing data at multiple time points and building a learning culture that enables both trainers and participants to shape the program's design and trajectory.

**Keywords:** Capacity Development, Monitoring and Evaluation, Learning, Gender, Agriculture, Research.

## Introduction

Integrating gender considerations into agricultural research and development priorities, implementation, and evaluation is critical to achieving gender equity and development goals (Meinzen-Dick et al. 2011). The United Nations promotes gender equality as a crucial priority in the Sustainable Development Goal 5, and global research has repeatedly underscored the significant human and economic cost of failing to equitably invest in the productive roles of men and women in agricultural systems (FAO 2011; Bryan et al. 2016; Njuki et al. 2016; Kristjanson et al. 2017; Gutierrez-Montes et al. 2020; Howland et al. 2021). However, efforts to achieve gender equality in agricultural research systems have been hampered by a lack of gender skills and capacities among researchers. One effort to address this barrier is to offer gender trainings to staff (EIGE 2016). Yet while more organizations are pursuing this, the quality standards of these trainings remain undefined, which ultimately limits the potential impacts of gender-related capacity development efforts (Bustelo et al. 2016).

Critical reviews of gender training further argue that the transformational potential of gender mainstreaming initiatives remains underutilized (Mukhopadhyay 2014) and that, in some cases, these initiatives serve to "instrumentalize" gender equality, further exacerbating and deepening gender inequalities rather than redressing them (Wilson 2015). A review by Mangheni et al. (2019), focused on gender training programs for agriculturalists in East Africa, reveals that gender courses do not sufficiently reflect on who should be trained, what the training content should include, or which methods should be used. Rather, the authors contend that many of these training programs have led to "churning out half-baked gender practitioners", without fostering any critical reflection on how gender inequalities are socially constructed (Sarapura Escobar and Puskur 2014) or on the role that researchers can play in reinforcing or challenging these inequalities. Despite this criticism, enhancing knowledge, skills, and behaviors for researchers to conduct in-depth gender analyses is critical to understand complexities and transform gender relations (Cole et al. 2014). As Njuki (2016) points out, efforts to develop capacities should not only consider building technical skills throughout the research cycle, but should also position researchers to better identify gender issues, improve understanding of the underlying causes of gender inequalities and encourage collaborative engagement with those affected to address them. In this sense, capacity development efforts require enhancing specific skills such as data collection and analysis techniques in addition to shifting mindsets and behaviors that ultimately influence the research produced by agriculture research systems and institutions.

Given the importance of fostering behavior change and the need for greater clarity on what constitutes an effective gender training program, a systematic and participatory monitoring and evaluation system that emphasizes learning needs to be embedded within gender training programs. This will contribute to a better understanding of how to effectively empower individuals as well as help capture the longer-term impacts of the training. Up to now, however, no such system has yet been proposed or tested. To fill this gap, we draw on our experience developing and implementing a monitoring, learning, and evaluation (MLE)[1] plan for the Gender-responsive Researchers Equipped for Agricultural Transformation (GREAT) course. In this paper, we also present lessons to inform future, effective MLE plans for gender trainings and further progress towards minimum quality standards for gender training programming.

Monitoring capacity development can provide critical insights for a project or intervention to improve its impact and sustainability. However, it can be difficult to operationalize, since it seeks to measure changes at multiple levels including in individual behavior and knowledge, as well as in organizational performance and the enabling environment (FAO n.d.). While capacity development[2] can play an important role empowering individuals and organizations to positively affect system change, it is widely agreed that such efforts should consider the different environments and circumstances people are in and how interventions such as trainings can be linked to a wider ecosystem of support (Nelson 2006; Asian Development Bank 2008; Pearson 2011). Varied individual and organizational circumstances also means efforts to effect change will occur in different forms, at different levels and over very different timescales (Mendizabal et al. 2011).

For a successful MLE plan, it is important to clarify the objectives of the training course and its envisioned contribution to medium- and longer-term outcomes. Understanding what is reasonable to measure at different time periods will influence the method and tools used during the data collection and analysis. A challenge to measurement is that the ways in which capacities develop after training are often not well understood; without sufficient evidence on change processes, incorrect assumptions may guide inappropriate methods (Horton et al. 2000; Taylor and Clarke 2008; Otoo el al. 2009). For example, evaluations of the impact of higher education show that many of the public benefits that education is expected to produce do not manifest themselves before 10-30 years (Davies 2012).

Unfortunately, in many capacity development initiatives, monitoring and evaluation primarily focus on short-term evidence of knowledge retention or immediate performance (Vallejo and When 2016). Linking capacity development solely to immediate changes in performance can fuel unrealistic expectations of short, direct paths between capacity development interventions and tangible productive results. Rather, "an understanding of capacity must also go beyond the instrumental, the technical and the functional and encompass the human, the emotional, the political, the cultural and the psychological" (Morgan 2006, 18). This is particularly true for advancing gender equality, since challenging entrenched gender norms can take time, is highly sensitive to individuals' experiences, and often depends on the presence of an institutional enabling environment (Guijt 2008; Hillenbrand et al. 2015). In this sense, it is useful to think about changes in capacity as in a "constant state of motion" (Ortiz and Taylor 2009, 87) whereby measurement approaches assess change as an incremental process which occurs at different points in time, rather than merely focusing on a final endpoint (Guijt 2008; Hillenbrand et al. 2015). Career paths and choices often progress in non-linear ways, horizontal movement is increasingly apparent, and progression can be towards different, inter or transdisciplinary fields. Therefore, measurement approaches should gather data at multiple time points and consider contextual factors at different stages.

## Contribution in favor of attribution

The unpredictable nature of capacity development also challenges efforts to determine attribution. A range of variables influence individual and organizational behavior change. In many cases, measurements of capacity development are not well suited to more quantitative and experimental approaches, which have long been associated with perceptions of tangibility and proof (Woolcock 2009; Mentz 2017).

Randomized control trials (RCTs), which estimate the mean net impact of an intervention by comparing results between a randomly assigned control group and an experimental group or groups (BetterEvaluation 2016), are regarded as the most credible source of evidence to guide decisions about an intervention's or project's effectiveness. However, capacity development rarely follows linear or monotonic trajectories and does not lend itself to randomization (Woolcock 2009; Mentz 2017). Instead, the focus is increasingly on "assembling evidence to show that one is learning diligently, adapting and taking a well-informed path" (Ortiz and Taylor 2009, 29) and less on "allocating credit for different levels of impact." Proponents of alternative approaches have drawn causal inferences on the basis of a combination of causes or a causal chain as opposed to a singular cause (Stern et al. 2012). In these approaches, the emphasis is on understanding the contribution of a particular intervention in a complex setting.

Theories of change, which are increasingly utilized in the context of complex development interventions, are a useful strategy and evaluation tool to identify what a program seeks to achieve, for whom, and through which pathways (Vogel 2012). Combining qualitative and participatory techniques with quantitative indicators can provide powerful insights on how change happens in tandem with descriptive measures and on what (if anything) has actually changed (Hillenbrand et al. 2015).

## Conceptual framework

In developing a MLE system, it is important to delineate the different levels at which changes are expected to take place. The Women's Empowerment Index framework, which is a context-specific composite index for the measurement of women's empowerment (Lombardini et al. 2017), defines three levels at which

change can take place: personal, relational, and environmental. While the more specific Women's Empowerment in Agriculture Index (Alkire et al. 2013) could also be used, the Women's Empowerment Index framework enables us to think more broadly about change in a research setting. In this framework, changes at the *personal* level refer to how individuals view themselves, their role in society, and their confidence in taking decisions and actions that have an impact on themselves and others. Applying this to a research setting, personal changes for agricultural researchers could be understood as shifts in how researchers view their roles and contributions to addressing gender inequalities, and their confidence and capacity to address them. Changes at the *relational* level take place in the relationships and power dynamics within an individual's surrounding network. Agricultural researchers are often highly interconnected, working with each other, households, and communities in order to design and implement their research. Thus, it is important to explore whether agricultural researchers are able to cross disciplinary boundaries, support one another, and effectively shape more inclusive and interdisciplinary forms of research within their research teams. Finally, changes at the *environmental* level refer to changes in the broader environment, including in social norms and attitudes. Here, the focus is on gathering evidence on whether changes among researchers and within research teams are contributing to any broader practice or policy related to the ways in which institutions conduct research.

Conceptually, the framework presented above highlights the crucial areas where the MLE system should explore changes. It is also built on the premise that empowerment is a multidimensional concept which is highly personal and context specific. This is particularly true for agricultural researchers who face very personal and professional biases depending on their roles, sex, types of agricultural systems, and institutional contexts in which they work. Given these diverse backgrounds and experiences, aspirations for training may significantly vary among the participants themselves and between the participants and those developing and implementing the training. Building a learning culture into the MLE approach whereby all participants have a voice and opportunity to shape its design and trajectory can help to develop more realistic and aligned training objectives among participants and implementers. It can also provide useful insights on what kinds of skills and capacities are needed to support individuals to advance gender-responsive research in their respective roles and institutions.

It is also important that the training itself, as well as the metrics and methods used to define its effectiveness, are inclusive and supportive of empowerment outcomes (SDC 2012). The use of participatory approaches enables participants to discuss issues that are relevant to their own needs and experiences, bringing greater authenticity, relevance, and collective ownership to the training objectives and evaluation (Estrella and Gaventa 1997). Both quantitative and qualitative measures are needed (Bamberger et al. 2010) and real-time feedback should be provided enabling capacity development programs to adjust to experiences and complex realities (Patton 2008). This makes MLE an integral part of capacity development interventions, both for those delivering programs as well as for participants. It ensures that capacity development objectives appropriately reflect the interests of both parties and cultivate a genuine partnership rather than a sense of "downward" development (INTRAC 2016).

## The evaluation context: the GREAT program and its MLE system

In order to capture how change occurs through capacity development, we present a MLE system developed in the GREAT project. GREAT is an innovative training program that seeks to equip agricultural researchers with theory, tools, and skills to move beyond "gender sensitization" to promote gender-responsive research[3] throughout the design, implementation, evaluation, and communication pathways. The GREAT model utilizes a mix of self-reflection, interdisciplinarity, applied learning, field application, and mentoring in order to create a comprehensive training package that promotes experiential learning (see Tufan et al., this issue).

While the GREAT model was refined and improved over successive cohorts, several core elements of the training approach were applied to all courses. These include:

i. Encouraging self-reflection of participants' own positionality and biases;
ii. Promoting appreciation and understanding across disciplines through team-based learning;
iii. Supporting applied learning through field application; and
iv. Nurturing an enabling environment through a targeted recruitment process that aimed to create a "critical mass" of fellows and the development of a community of practice which provided access to resources, networking, and connections across courses and institutions.

Over five years (2015-2020), GREAT delivered five open enrolment courses.[4] As the demand for GREAT training increased at institutional and project levels, experimental customized "spin-off" courses of three to six days of technical instruction adapted from the original model were developed collaboratively. GREAT trained 292 researchers from 31 countries over five years, from multiple institutions with various roles and responsibilities, and including a wide range of interdisciplinary research themes.

### *Designing the GREAT MLE system*

The GREAT MLE system was designed to provide both real-time and long-term insights to a range of stakeholders. Understanding the value of the different components of the training, delivery methods, and the extent to which participants' needs and expectations were met was critical to continuously improving the course. In particular, it was important to capture how novel features of the GREAT training, such as the fieldwork component, mixed team-based approach, and efforts to challenge existing biases and positionality, were perceived and valued by participants alongside an appraisal of training content, tools, and

delivery. We developed the overall MLE system, tool preparation, data collection and led the analysis in consultation with the GREAT program team.

A utilization-focused approach to MLE was applied at the beginning of GREAT, rooted in the key principle that the evaluation of the program should be useful to its primary audiences (INTRAC 2017b). The primary audiences and intended uses are as follows:

i. *For the GREAT program team*, to inform changes in course delivery and design, ensure accountability, and demonstrate impact.
ii. *For participants*, to support improvements in their gender-responsive research methods and approaches and understand barriers and enablers for applying learning.
iii. *For organizations*, to demonstrate the value of their investments and encourage further institutional support for gender-responsive research.

The MLE system used a mixed method, theory-based approach (INTRAC 2017a). This approach was applied for several reasons. The research pool interested in the GREAT program are highly interconnected and many of the prospective participants collaborate on research articles and influence one another through advocacy, mentoring, and information sharing. For this reason, establishing an appropriate comparison group of a sufficient size while eliminating spillover effects was neither a realistic option nor would it have provided detailed insights on what was driving change amongst a highly diverse group.

Instead, the MLE system was designed to capture a range of aspects of the training design, implementation, and progress towards the short- and medium-term outcomes and the long-term impacts. In the short term (during the course), evidence was sought on whether a high-quality training was being delivered and whether there were perceived shifts in fellows' competencies for gender-responsive research.[5] In the medium term (6-12 months), the focus was on understanding the extent to which learning imparted through the course was translated into practice in research design, planning, implementation, and communication. A second goal was to understand whether post-course training support and resources were relevant and useful. For the longer term (3-4 years), the MLE system sought to capture whether the changes reported in gender-responsive agricultural research reflected good practices, whether there were patterns in reported changes over time, and how contextual factors were driving or limiting application in very different research environments.

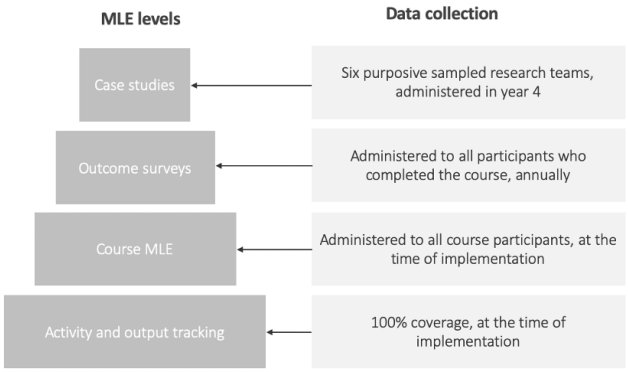


**Figure 1** Overview of data collection levels.

Figure 1 presents an overview of the different layers of the MLE system, while Table 1 sets out a detailed overview of the timeline, different methods, and rationale behind the data collection processes.

While efforts to measure and quantify women's empowerment are well-documented (Hillenbrand et al. 2015; Lombardini et al. 2017; Alkire et al. 2013) and much has been written on the different methods used to assess capacity development initiatives (see, for instance, Preskill and Boyle 2008; Blume et al. 2010), the literature on their intersection with agricultural research settings is nascent. This paper addresses this intersection, and draws out key lessons on the importance of integrating participants' voices, collecting data consistently over time, and using different qualitative and quantitative methods. It aims to showcase examples of the different tools used and results achieved within the context of the GREAT training program, underscoring which factors worked and did not work.

## Methods

For this paper, the authors, as the MLE team, draw on various resources compiled during the five-year period of the GREAT courses (2015-2021), in addition to the data collected through the MLE system. These resources include: participant observations during GREAT open-enrolment courses and selected customized courses; documents detailing the methods, tools, and approaches; and a synthesis of the annual reflection sessions from the MLE and program team. We also present a subset of the MLE results to highlight key learnings and approaches. A detailed discussion and appraisal of the results of the GREAT program are presented elsewhere (see Tufan et. al., this issue).

## Data collection

MLE data were collected before, during, and annually for multiple years after the course from researchers, fellows, trainers, and senior leadership from participating institutions. These data allowed for analyses that could explore the complexity of capacity development over time, as well as at different levels.

### *Data collection before the course*

Prior to the start of the first course, researchers, trainers, and senior leadership from targeted research institutions engaged in a MLE workshop to provide input on the training approach and expected outcomes. Using an actor mapping process (FSG 2015), participants defined key organizations and/or individuals that

| Timing/Focus | Method(s) | Approach | Rationale |
|---|---|---|---|
| **Pre-course** *Prior to the course start* | Training application; pre-training survey | Data were collected from prospective training teams through an online training application form. The application solicited information on fellows' existing research, reasons for participation, previous experience in gender training and field data collection. All accepted applicants completed an online pre-training survey which gathered biographical data and included a pre-training technical assessment. | Capturing baseline data on participants' motivations, funding, research details, pre-training expectations, experiences, and pre-training knowledge. |
| **During course (fellows)** *Continuous* | Course evaluations; key informant interviews; course observations; learning rubric | Data were collected through two course evaluations, key informant interviews, direct observations, and a learning rubric. At the end of week one and week two of the training, all fellows were asked to complete course evaluation forms to provide feedback on their learning and experience with the training course. Semi-structured key informant interviews were held with a cross-section of participants. Half of the selected informants were interviewed again in week two in order to capture changes in perspective. | Capturing feedback on what was working well, what was not working well, and areas of improvement going forward to direct near real-time changes in course implementation; assessing participation and learning dynamics; gathering voices and experiences throughout the course; and assessing changes in attitudes, knowledge and skills, and competencies of the fellows. |
| **During course (trainers)** *Continuous* | Observations; assessment rubrics; discussion group | Debriefs by trainers occurred at the end of each course session day. During the first course (2016-2017), the entire training team met at the end of each day for lengthy discussions (1-2 hours) of the day's sessions. For courses 2-4, trainers used sticky notes to track pros, cons, and changes in each session, with a brief review (10 minutes) of all comments. Trainers' debriefs were conducted after the end of each course. Participants reflected and discussed each session as well as general aspects of the model based on the compiled comments and suggestions from the daily briefings during the course. | Enabling near real-time learning and improvement of course implementation and capacity development for the trainers; capturing views from the training team; and documenting learning after each training for continual improvement. |
| **At least 6 months after the course ended (fellows)** *Annually* | Online survey; key informant interviews; document review | Administered to all training participants who completed the full training course. Data were collected through an online survey (i.e., Survey Monkey), using both open-ended and closed questions. Careful attention was paid to qualitative question design in order to ensure that the collected data were verifiable and consisted not only of fellows' perceptions, but they also elicited detailed information with specific examples of change to contextualize the quantitative responses. Follow-up interviews and requests of documents were conducted where responses were incomplete. | Assessing how the change in attitudes, knowledge, skills, and abilities is translated into more effective gender-responsive research practices in project and institutional level outcomes (collecting data annually increased sample sizes and enabled analysis of research application over time). |
| **Year four (fellows)** *Once* | Key informant interviews; focus group discussions; observations during field visits; document review | Six in-depth purposive sampled case studies selected from participating teams in course 1 (Root, tubers, and bananas (RTB)) and course 2 (Cereals) to ensure sufficient time for learnings to be applied, and possible benefits to materialize or to be analyzed and interpreted. Through the case studies, data were collected through key informant interviews with fellows, other research colleagues, supervisors, implementing partners and farmers. These were complemented with focus group discussions, observations during field visits, and extensive reviews of documents. | The case studies analyzed the accomplishments of the teams by examining their GREAT-associated research. However, case studies also reviewed subsequent research materials shared with the evaluation team and considered individual impacts as well as explored dynamics between cohorts (i.e., the impact of multiple teams trained from the same institutes or programs). Assessing quality and type of changes reported over time and exploring contributing factors to practice change. |
| **On-going (project)** *Part of project monitoring* | Project MLE system | Project database developed to track attendance in GREAT-linked events, journal article submissions, requests for technical assistance, financial data, resource downloads, and click-through rates. | Enabling triangulation of information; providing data points on different levels of engagement and participation with GREAT. |

**Table 1** Overview of timing, methods, and rationale for data collection in GREAT.

comprise the agricultural research system and the outcomes to which they contribute. They also identified the enabling and constraining factors, as well as the underlying assumptions about what needed to be in place in order for change to happen, see Table 2. This approach places people at the heart of each change and more closely reflects real-world contexts as compared to linear result frameworks or linear theories of change. This process of co-construction also enabled the program team and workshop participants to challenge assumptions about how change happens and to broaden the range of strategic options for design, implementation, and measurement of the training course.

Before each course, applicants also provided individual information, including biographical data, levels of experience with gender-responsive data collection and analysis, research role, and interest in the course. Applicants were required to apply as interdisciplinary teams of two or three and provide information on: how they thought that the training would help them to advance their work; the research project that would be used for the course;

| Actor | Actor-level outcomes | Enabling factors | Constraining factors | Assumptions |
|---|---|---|---|---|
| **Trainers** | O1: The trainer needs to deliver a strong curriculum that motivates, builds capacity, and strengthens the performance of GREAT participants. There must be a strong encouragement to complete the course completion, with focus on the implementation and analysis stages. | Availability of bank of training resources to help to drive an innovative curriculum.<br><br>Regular contact with some members of GREAT leadership to share questions and discuss direction.<br><br>Several GREAT partners that can aid in different technical areas and can brief trainers appropriately.<br><br>Engaging pedagogy advocated by the curriculum designers and a "sandwich" approach help to keep up the engagement of fellows. | Lack of contact with core GREAT team.<br><br>Ingrained models of rote learning.<br><br>Narrow range of backgrounds/interests. | Trainers are well qualified and available to do the job.<br><br>Trainers are consistent in their approach.<br><br>Trainers have been fully briefed on the vision, desired outcomes, and activities of the GREAT program. |
| **GREAT Fellow** | O1: GREAT Fellows need to actively engage with the course content, apply it, and be open to using it as a framework to influence their future research.<br><br>O2: GREAT Fellows need to have had an attitudinal change in how they view gender-responsive research and use this change to motivate them to influence others to adopt this approach<br><br>O3: GREAT Fellows respond to the MLE system and implement MLE as a critical part of their future research to begin to generate an evidence base for the impact of gender-responsive research. | Applied and secured a place through a competitive process, demonstrating a keen enthusiasm for further awareness of gender-responsive research.<br><br>Availability of funds.<br><br>Donors flexible to funding changes mid-course through re-structuring of target trackers.<br><br>Institutional support for gender responsiveness.<br><br>Close working relationship with gender researchers. | Lack of understanding flexibility by mangers/donors.<br><br>No enabling environment for gender-responsive research.<br><br>No enabling environment for gender-responsive research.<br><br>Competing research priorities. | Fellows have the funding and institutional support to change approaches and methods.<br><br>Training is effective and equips researchers to change approaches/methods. |

**Table 2** Selection of actors' descriptions, outcomes, enabling and constraining factors, and assumptions.

and the ways in which gender considerations were reflected in their work. For selected candidates (i.e., fellows), a needs assessment and pre-course technical assessment were also administered through an online survey. This information provided important baseline data of incoming expectations, participant experience, and starting points.

### *Data collection during the course*

Information was collected simultaneously by the external MLE team as well as internally by trainers, mentors, and the program management team. Several methods were used: pre- and post-assessments, independent observation, participant feedback, daily team reflections, notes on improvements in research design carried out during the course, end-of-course evaluations, and end-of-course debriefs by trainers.

The external MLE team engaged in routine informal and confidential discussions with fellows, organizers, trainers, and partners throughout the training. This enabled the MLE team to gather participants' voices and nuanced insights on fellows' experience during the first week of their training, the four-month fieldwork, and their return for the second and final week of training. This was complemented with independent observations whereby session quality and level of engagement were systematically documented by a member of the MLE team through a structured checklist.

Course evaluations were designed to capture self-reported competencies, most valuable skills gained, change in perceptions of the value and importance of gender-responsive research (henceforth, GRR), value of the course, satisfaction with the course design, sessions, training content, as well as trainers' competency and delivery methods. Two course evaluations were administered, one after the first week of the training and one after the final week. This enabled tracking of how perceptions changed, particularly after the exposure to different components of the training, such as gender-responsive farmer level data collection, which for many was an entirely new experience.

As part of the evaluation, fellows rated their proficiency across 17 technical competencies covering a broad range of topics, from the course curriculum to the competencies intended to support application post-course. These ranged from competencies on technical concepts – such as the ability to recognize and avoid using gender stereotypes and the ability to use mixed methods to collect data – to more tactical skills – such as the ability to identify and source relevant gender expertise. These were then mapped against the training's five core objectives of (1) gender concepts and principles; (2) appreciation and value; (3) design and planning; (4) collection and analysis; and (5) communication.

Where possible, the tool remained the same over the years in order to enable comparison across themes. However, some training sessions were added while others were dropped, and training

| Year | Theme(s) | Total Response | F | M | Response Rate | Tools |
|---|---|---|---|---|---|---|
| 2017 | RTB | 21 | 14 | 7 | 66% | Survey |
| 2018 | RTB, Cereals | 52 | 27 | 25 | 87.7% | Survey, KIIs |
| 2019 | RTB, Cereals | 24 | 11 | 13 | 59% | Survey |
| 2019 | Legumes | 16 | 6 | 10 | 76% | Survey |
| 2020 | RTB, Cereals, Legumes | 43 | 10 | 33 | 60% | Survey |
| 2020 | Plant breeding | 16 | 5 | 11 | 62% | Survey |
| 2020 | 4 Customized courses | 52 | 24 | 28 | 51% | Survey, KIIs |

**Table 3** Overview of outcome survey response rates and tools.

objectives were not formalized until after the first theme. This meant that some measures, for example the measure of attitudinal changes, varied from the first theme compared to the remaining ones.

### *Data collection annually post-course*

Surveys were completed for each cohort every year after the first training course in 2016-2017. Fellows reported the actions that they took to integrate GRR approaches into their existing research and new proposals, as well as whether they had produced any gender-responsive communication products. Information was also gathered on whether fellows had taken any action to support changes within their research institutions, whether any institutional change was triggered because of these actions, and whether they faced any barrier in their efforts to advance GRR.

Table 3 provides an overview of the data collection instruments used and annual response rates. One year after the completion of the course, fellows were also asked about the value of post-training resources and their participation in the GREAT community of practice to understand the extent to which post-course services were useful in generating new connections. Fellows were initially given the same survey regardless of when they had completed the course; however, a pared-down survey was administered in 2019 to earlier cohorts that did not ask about post-course services.

Both quantitative and qualitative data were collected through outcome surveys. These captured information on standardized metrics and more open-ended perspectives. For example, fellows were not asked to describe all the changes they made in terms of GRR, but to report the most significant changes they contributed to and to provide concrete examples illustrating these changes. The most significant change (MSC) technique is an established form of participatory monitoring and evaluation (ODI 2009). It involves the regular collection of significant change stories from program beneficiaries followed by a systematic selection of the most important ones by designated stakeholders or staff (ODI 2009). Within GREAT, responses were not systematically filtered, but rather the focus was on collecting a range of experiences and accomplishments that researchers deemed important. This avoided pre-defined measures of success. Qualitative questions were also designed to ensure that the data were verifiable. They consisted not only of fellows' perceptions, but they also included detailed information with specific examples of change. These qualitative data were used to complement the more standardized, quantitative responses. A sample question of the outcome monitoring is presented in Box 1.

> *Please describe in more detail the most significant change(s) that you have been able to apply to your research after the GREAT course. Please use specific examples to illustrate the change(s) (e.g., name of the research project, concepts used, the type of activities you undertook, types of methods and/or tools used; guidelines used, etc.) to show how what you do now differs from your previous approach. Provide a link for each of the projects if available.*

**Box 1** Sample qualitative question of the outcome monitoring survey.

Quantitative questions asked fellows to rate the level of application of gender-responsive methods and approaches to their own research, see Table 4. The retrospective baselines approach was used to establish the level of application before the fellows' partic-

| | Very low | Low | Moderate | High | Very high |
|---|---|---|---|---|---|
| Before participating in the GREAT course | □ | □ | □ | □ | □ |
| 1 year ago | □ | □ | □ | □ | □ |
| At the time of the survey | □ | □ | □ | □ | □ |

**Table 4** Sample quantitative question of the outcome monitoring survey.

ipation in GREAT. Retrospective baseline design is considered a convenient, valid method for measuring self-reported change (Klatt and Taylor-Powell 2005). This approach allows for a single point of administration, enables respondents to reflect on their experiences at one point in time, and helps to avoid response shift effect. The latter occurs when a respondent's frame of reference changes significantly during the course of the intervention (Lam and Bengo 2003; Mentz 2017). To establish the extent of change in the application of gender-responsive methods, the proportion of fellows reporting different levels of application was compared before and after participation in the GREAT course. The time elapsed was also used to ascertain possible trends. Quantitative data were disaggregated by cohort, paying particular attention to differences between groups, such as between men and women, and seniority and disciplines of researchers. Researchers' disciplines were broadly classified as either social science or biophysical science.

Fellows were then asked to report on the factors and/or conditions that enabled them to apply (directly or indirectly) the knowledge and skills acquired from the GREAT training to their work. This may be related to their personal characteristics, such as being in a position of authority, or as a result of a more conducive research environment, such as having other colleagues that had been trained in gender-responsive research.

Incomplete information provided in the online surveys was clarified through follow-up emails and telephone interviews. This was an important, albeit time-intensive step which led to the collection of better-quality data that could ultimately be verified (e.g., accurate names of journal submissions or research links, etc.). Information was then compiled into annual outcome reports and data were presented to the training teams in annual reflection meetings in order to facilitate discussions on progress towards learning objectives and to identify the additional evidence needed.

### *Case studies to assess changes over multiple years*

In the fourth year, a series of independent case studies were carried out to assess the quality of GRR and to understand what factors were driving or constraining practice changes. Case studies are a powerful tool to learn about change in complex environments, since they draw on multiple years of data and sources, are more comprehensive than traditional surveys, allow for flexibility, and enable the space to search for alternative explanations for the observed changes (Balbach 1999).

A critical component of the case study methodology is defining the case or "unit of analysis". According to Miles and Huberman (1994), the case can be understood as "a phenomenon of some sort occurring in a bounded context" (Balbach 1999). In the case of GREAT, the unit of analysis used was the team. This enabled an exploration of the team-based training model. We also examined relational change (i.e., intra-team dynamics) and the extent to which this can influence broader environmental change. The case studies provided a link to the initial research submitted for the course and to any work that was directly informed by the field research carried out as part of GREAT. The expected and reported changes cited in field reports/presentations, informational interviews, and the outcome surveys enabled us to explore achievements outside of the training environment.

Case studies examined if project teams had amended their existing research projects to reflect a more gender-responsive approach, verified the level and quality of gender-responsive practices applied, and analyzed what combination of factors had enabled project teams to integrate GRR approaches into their projects. The core evaluation questions applied to the case studies are presented in Box 2.

Seven teams[6] were originally selected as case studies, from the first- and the second-year training cohorts. The intention was to ensure sufficient time for learning to be applied and for possible benefits to materialize for analysis and interpretation. Both higher and lower performing teams were selected to help to identify contributing factors in different contexts. Performance was determined by the quality and improvements made in their GRR proposals during the course and the level of practice changes reported in the annual outcome surveys. Selection criteria for the case studies varied slightly from year one to year two based on changes in the course design.

Data on the country's context, institutional history, and policies related to GRR were collected through document reviews and key informant interviews. An understanding of fellows' roles and contributions to GRR practices was developed from the annual outcome monitoring reports and review of referenced project documents and tools, and by talking to colleagues and supervisors. Where possible, field visits were carried out to inter-

1. Have GREAT fellow teams applied high quality gender-responsive research practices to their research projects?
2. If so, what led to the effective implementation of gender-responsive research practices within their research project? What is the evidence that GREAT is a key contributor to these changes? Are these changes of sufficient quality and leading to positive outcomes for project beneficiaries and/or gender-responsive research?
3. Have interdisciplinary teams worked together to conduct gender-responsive research? If yes, how has this manifested in their GREAT-linked projects, within their institutions or additional projects they have been involved in?
4. If not, why did the training not translate into practical changes within the project? What other factors may have prevented or limited the application of gender-responsive research practices to their research projects?

**Box 2** Case study evaluation questions.

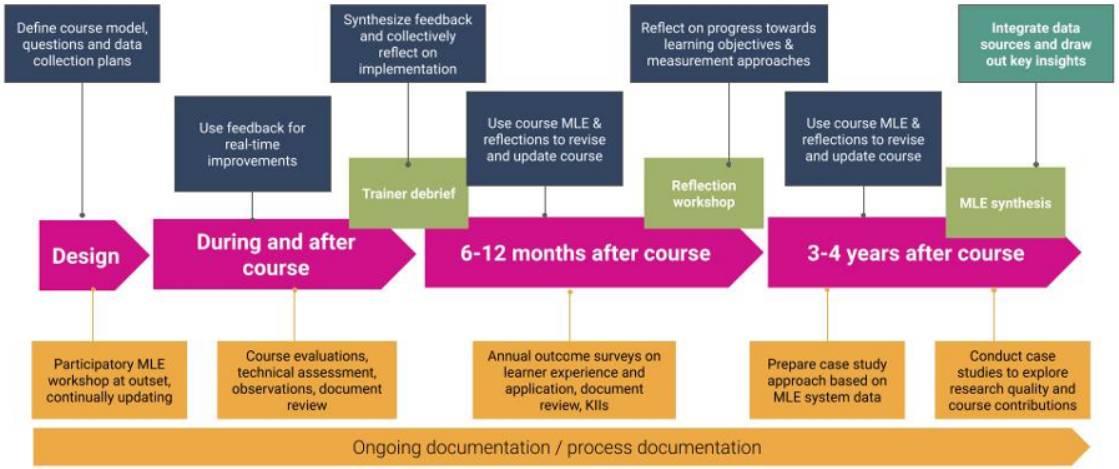


**Figure 2** Overview of the MLE system process.

view implementing partners and conduct focus group discussions with farmers who engaged with the fellows in order to capture multiple perspectives on how GRR approaches had been applied.

## Data analysis, results, and use

An overview of the different stages in which the data were collected, analyzed, and used is presented in Figure 2. This breaks down the system into key stages including design, during and immediately after the course, 6-12 months after the course, and 3-4 years after completion of the course.

Quantitative data from course evaluations and outcome surveys were analyzed using SPSS and simple descriptive statistics (e.g., mean and percentages) were used to present the findings. Using ordinal scales, the course evaluation illustrated the ratings by fellows of self-reported changes in attitude, knowledge, skills, and competencies targeted by GREAT. Given the small sample sizes, weighted averages were computed and used to give a better representation and ranking of issues using a Likert scale. Qualitative data from course evaluations and outcome surveys were used to contextualize and verify quantitative responses. Thematic content analysis (Nowell et al. 2017) using both deductive and inductive inquiry was applied to identify patterns and themes.

### *Course relevance and effectiveness*

Regularly testing how different features were valued and experienced by fellows, analyzing feedback, and encouraging regular input on how the course could be improved was central to building a better picture of what training methods were effective, for whom, and why.

As an example, Figure 3 presents a breakdown of the course feature ratings by seniority from the first course. Senior and mid-level[7] respondents indicated slightly higher levels of satisfaction with the technical content than junior fellows. However, the reverse was seen with regard to field research, with senior participants rating this lower than mid-level and junior participants.[8]

Data were then triangulated with key informant interviews to identify the key factors contributing to these trends. Several respondents noted that the field research experience was illuminating, particularly for those who would regularly send colleagues to the field but had not previously gone themselves. Field research itself was very time- and labor-intensive for fellows in senior positions. Refinements were implemented in the field research process, such as opting for a more gradual approach where data were first collected and then jointly analyzed. This led to overall improvements in satisfaction across training cohorts.

Another example of how feedback was purposefully integrated comes from the course evaluation form. Fellows were asked to indicate their level of satisfaction with the sessions covered

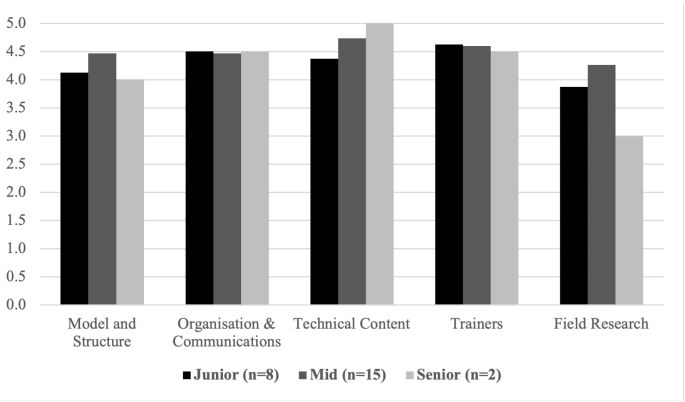


**Figure 3** Fellows' reported satisfaction with different elements of the GREAT course by management authority. *(Source: GREAT course evaluation, course 1.)*

| Session not satisfied with | Area of concern | Recommendations |
|---|---|---|
| Quantitative data analysis principles and practice. | • The content was too basic for those with quantitative training.<br>• Case studies were not well explained. | • Focus on main principles, e.g., interaction terms vs separate regressions.<br>• Make the session optional for the breeders and engage with the quantitative researchers only.<br>• Use clear case studies. |

**Table 5** Areas of concern and recommendations of fellows. *(Source: GREAT evaluation report, theme 4, week 2.)*

in each week (using the following scale: 4=extremely 3=satisfied, 2=partly satisfied, and 1=not satisfied at all). Fellows were asked to assess each session on the basis of its content, delivery methods, and added value to themselves and to their work. For the sessions with which they were only partly satisfied (rated 2 or below), fellows were asked to describe their areas of concern and to offer recommendations for improvement. This feedback was then synthesized and shared with the training team during debriefs by trainers and curriculum review meetings. Table 5 provides an example of the kind of concerns raised, the rationale behind these concerns, and the proposed recommendations.

The benefits of unique course features were also explored over time. For example, the value of the team-based approach to training was of particular interest for the GREAT program. This was assessed at several points in time: during the course, and specifically during field research through independent observations and reports by mentors[9]; and immediately after the course through key informant interviews. Fellows of different seniority, disciplines, and from different categories of institutions were asked about their team experiences. The use of thematic content analysis helped to identify commonalities such as increased appreciation for interdisciplinary perspectives. The data were triangulated with discrete observations from team mentors who could provide more context on some of the key challenges, such as unequal distribution of work and issues with coordination which were exacerbated when teams were working in different locations. Complementing team-based accounts with more independent observations from mentors was important to counterbalance possible biases from peer assessments.

Perceptions of the utility of the team-based approach were then compared with feedback from key informant interviews, which were carried out several years after participation in GREAT courses, in order to assess whether initial accounts of the potential benefits of team-based training actually materialized. Accordingly, 12 out of 32 key informants participating in the customized courses discussed having colleagues and supervisors who were aware of and exposed to GRR concepts as one of the most critical enabling factors to apply what they learned in the course. The framing of the question around the factors that contributed to the application of learning allowed participants to come up with their own responses, rather than explicitly asking them to reflect on specific features of the course. A sample quote is provided below to illustrate the kind of information that was provided by respondents.

> *Having a supervisor who was exposed to GRR and kept reminding me about the things we talked about during the course. If not for this, I would be doing the normal qualitative data collection but not gender-responsive qualitative research.*
>
> (Key informant interview, female social scientist, 2020)

Building on the feedback which was shared at multiple points in time by those delivering and participating in the course, coupled with an appraisal of progress towards key learning objectives (discussed below), was central to confirm the suitability and effectiveness of the course model.

### *Researchers' attitudes and competencies during and immediately after the course (course MLE)*

Capturing changes in attitudes is notoriously challenging and often highly dependent on personal circumstances. However, it remains an important domain to explore as behavior change depends not only on competence, but also on the willingness to apply that competence (ten Cate and De Haes 2000). Moreover, the course had an explicit objective of encouraging researchers to better understand and challenge processes and approaches that reinforce inequalities. Evidence of attitudinal changes was gathered through several methods. Fellows were asked to rate the extent to which the course had changed their attitude towards GRR (from "not at all" to "significantly") after the first week and at the end of the second week. They were encouraged to provide specific examples of what they used to do before and of which elements of the course contributed to changing their approaches to research. Results from the first week were compared to those of the second week in order to understand whether additional features of the course, such as the field research component, were important factors in shifting perceptions. Qualitative data collected through key informant interviews and course evaluations were particularly important for contextualizing quantitative responses. The interview excerpt below illustrates how individuals were engaging in the self-reflection process and considering how research processes can impact the utility and benefit of end users.

> *What we just realized is that all people that have been trained as seed interpreters are men, most of the participating farmers are men… So, we are now thinking how are these people identified. The criteria were simple, you must have 10 acres or 5 of land, but clearly can you can see this puts the women out of the picture… So, we want a total*

| | Women | | Men | |
|---|---|---|---|---|
| **Technical competencies** | **Bio** | **Social** | **Bio** | **Social** |
| Use gender concepts and principles correctly | 3.7 | 4.0 | 3.6 | 4.0 |
| Identify gender-based inequalities, constraints, and opportunities | 3.7 | 4.3 | 3.7 | 3.8 |
| Recognize and avoid the use of gender stereotypes | 3.7 | 4.2 | 3.9 | 4.3 |
| Appreciate the usefulness of gender research in your work | 3.7 | 4.4 | 4.4 | 4.5 |
| Feel motivated and self-driven to integrate gender into your work | 3.3 | 4.4 | 4.4 | 4.5 |
| Assess your organization to identify gaps and actions | 3.0 | 3.7 | 4.0 | 3.7 |
| Formulate a gender research question | 3.7 | 4.4 | 4.0 | 4.5 |
| Support the gender research question with relevant theory | 3.3 | 4.1 | 3.7 | 4.0 |
| Identify and source relevant gender expertise | 3.7 | 4.3 | 4.3 | 4.3 |
| Develop data collection tools | 3.3 | 4.3 | 3.6 | 4.5 |
| Develop sampling framework | 3.3 | 4.1 | 3.4 | 4.3 |
| Use mixed methods to collect data | 3.7 | 4.3 | 4.0 | 4.3 |
| Use mixed methods to analyze data | 3.3 | 4.2 | 3.3 | 4.3 |
| Use mixed methods to interpret and integrate findings | 3.5 | 3.9 | 3.3 | 4.0 |
| Package research outputs for different audiences | 2.7 | 4.0 | 3.7 | 3.8 |
| Communicate outputs of research in a gender-responsive manner | 2.7 | 4.0 | 4.0 | 3.8 |
| Share learning from the GREAT course with peers and leaders | 3.3 | 4.1 | 4.3 | 4.3 |
| Sample size (n) | 3 | 9 | 7 | 4 |

**Table 6** Distribution of fellows' self-reported proficiency in GREAT key themes. *(Source: Course evaluation results, course 2.)*

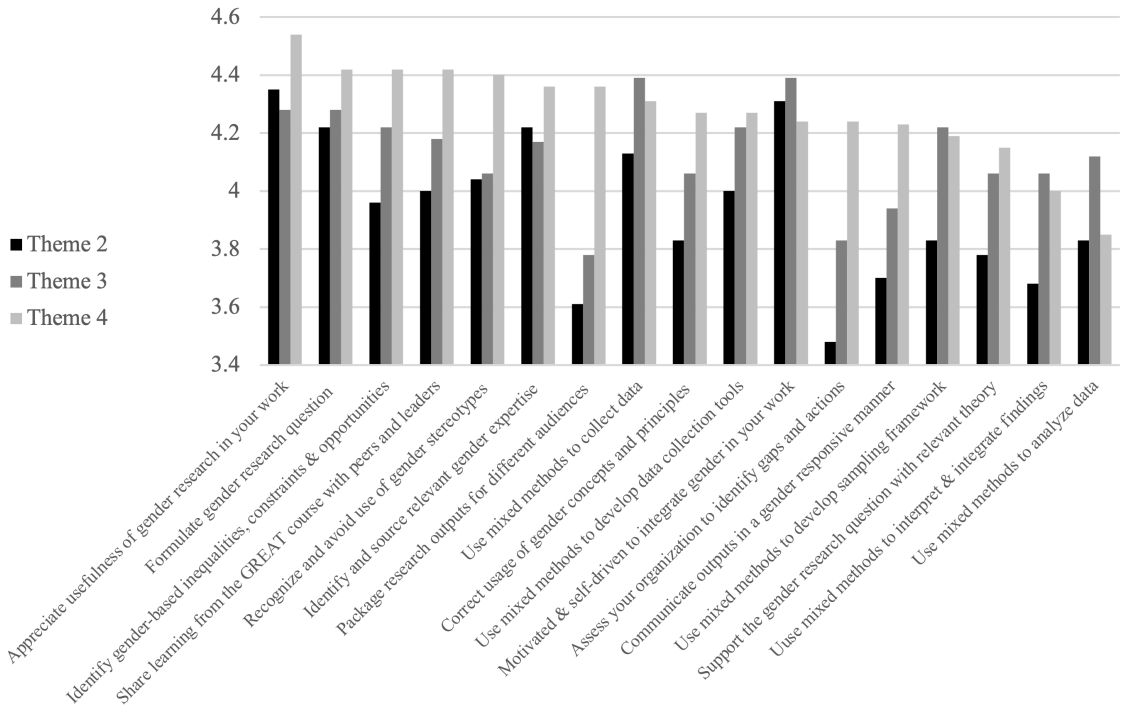


**Figure 4** Reported competencies across course themes. *(Source: Course evaluation reports, courses 2-4.)*

*change in our operational procedure…If we invite people for training, we want to be specific, deliberate[ly] selecting who is this going to consist of.*

(Key informant interview, male crop breeder, 2017, Uganda)

Disaggregating data by discipline provided useful insights on factors influencing attitudinal changes. Interestingly, even gender specialists (although notably self-ascribed) reported learning basic gender concepts, which in some cases was attributed to changing their outlook. This was important to understand in a course that trained people with very different technical backgrounds as it emphasized the need to build a common understanding of gender concepts regardless of technical training.

*Before coming to this course, my thinking was sex is equivalent to kind of gender… So, to me this training was more of an eye opener to try and understand this whole concept of gender, to be able to understand the difference between sex and gender, to know that gender goes beyond just knowing this is female, to understand the different dimensions of the whole concept, the kind of relations, the kind of constraints, the kind of opportunities, the kind of roles, resource access and control – a broad spectrum of what gender means. So, to me my experience through this GREAT course has been quite enriching I came as a blind person but I can say that now I can see a green light ahead.*

(Key informant interview, female gender specialist, 2017)

Reviewing technical competencies highlighted where proficiencies were perceived to be stronger or weaker, while disaggregation by gender and discipline allowed to thoroughly examine patterns across research characteristics. Table 6, from the second course, shows reported proficiencies presented by gender and discipline. In this cohort, differences were observed across disciplines as well as between genders, with female biophysical scientists reporting lower competencies across all but two categories.

Figure 4 reveals several patterns over time, including where there were significant differences in reported levels of competencies between courses and where there were consistently higher or consistently lower proficiencies. For example, "ability to appreciate the usefulness of gender focused research in their work" and "[felt] motivated and self-driven to integrate gender into their research work" were consistently scored higher than other competencies. Fellows' reported competency on the use of mixed methods to collect and analyze data were comparatively lower, suggesting an appreciation of the difficulties involved in becoming proficient in this skill set and the need for further technical instruction.

The combination of objective observations of course delivery, technical and perception-based assessments, and qualitative interviews provided important insights on course learning and experiences among diverse participants.

### *Application of knowledge, skills, and approaches learned (outcome monitoring)*

Qualitative and quantitative responses in the annual outcome monitoring were analyzed to understand which fellows reported applying GRR methods after completing the course. Fellows were asked to rate their application on a scale from very low to very high. Figure 5 shows fellows' accounts of how application had varied over time, revealing that between 65 to 79 percent of fellows indicated high or very high levels since 2017. This is contrasted with the 61 percent of fellows who rated their application as low or very low before GREAT. This was important for the MLE system as it provided an overview of how application changed over time. Notably, most fellows reported a large increase in their level of application the year after their participation in the course, with fairly consistent application in the following years. This points to an initial momentum linked to the participation in the course, but it also suggests consistent application years after the course ended.

Data were also disaggregated by gender and discipline to explore whether there were differences in the reported application. Figure 6 shows slightly higher levels of application among social scientists, with 77.8 percent of social scientists reporting high or very high application one year after the course compared

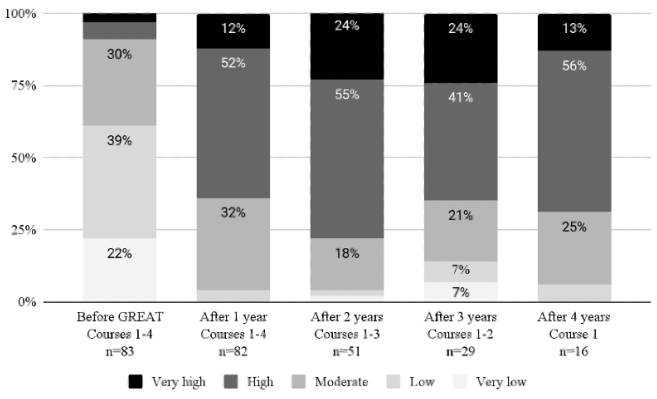


**Figure 5** Fellows' self-reported application of gender-responsive research over time. *(Source: Annual outcome surveys, 2018-2020.)*

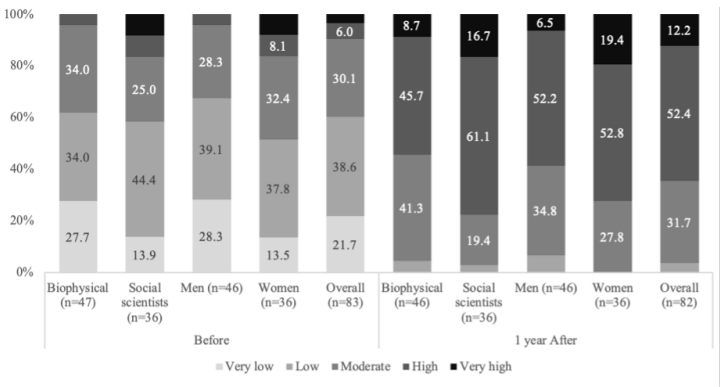


**Figure 6** Fellows' self-reported application of gender-responsive research by sex and discipline. *(Source: Annual outcome surveys, 2018-2020.)*

| Key themes | 2017 (RTB) | 2018 (RTB, Cereals) | 2019 (RTB, Cereals, Legumes) | 2020 (RTB, Cereals, Legumes, Plant breeding) |
|---|---|---|---|---|
| **Awareness-raising collegial, social circle or community level** *(training colleagues, partners, sharing learning)* | 0 | 5 | 4 | 2 |
| **Attitudinal shift and knowledge acquisition** (*cited increased appreciation, importance, and other knowledge gained*) | 3 | 3 | 0 | 1 |
| **Gender-responsive research design and planning** (*explicit focus of gender in research proposal and design*) | 1 | 1 | 4 | 4 |
| **Team building** *(greater involvement of disciplines, gender experts, and demand for trained fellows)* | 5 | 0 | 2 | 3 |
| **Qualitative data collection** *(use of KIIs, FGDs, applying skills on interviewing, facilitating, and capturing detailed info)* | 0 | 11 | 3 | 7 |
| **Mixed method data collection, analysis, and reporting** *(qualitative data collection with surveys, integrated into reports)* | 0 | 14 | 8 | 9 |
| **Collection, analysis, and/or reporting of sex-disaggregated data** *(shift from general focus groups to capture sex and age data)* | 6 | 13 | 4 | 6 |
| **Gender-responsive qualitative data analysis** *(applying different methods of qualitative data analysis)* | 6 | 3 | 0 | 2 |
| **Gender-responsive research approaches in implementation of biophysical research** *(deliberate targeting of men and women in trials, activities, and evaluations)* | 2 | 13 | 3 | 17 |
| **Use of social science to analyze gender-responsive research findings and inform biophysical research interventions** *(design of trials, treatments, breeding objectives)* | 0 | 6 | 1 | 9 |
| **Adoption of guidelines or gender-responsive operational policies** *(use or development in research or work environment)* | 0 | 2 | 2 | 1 |
| **Provision of budget to implement gender-responsive research** *(earmarking funds or resources for* gender-responsive research*)* | 0 | 1 | 0 | 0 |
| **Gender-responsive dissemination strategy** *(use of gender-responsive data to inform product marketing and dissemination)* | 0 | 0 | 1 | 1 |
| **Gender-responsive evaluation** *(impact of interventions on gender, men and women represented in the evaluation)* | 0 | 0 | 2 | 1 |
| **Gender balance in educational opportunities** *(ensuring that both men and women are represented in educational advancement opportunities)* | 0 | 0 | 2 | 1 |

**Table 7** Codes for significant change stories from fellows' outcome monitoring. Gray shading reflects a heat map for each year. The darker the gray, the more frequently the theme occurred in that year.

to 54.4 percent of biophysical scientists. Similarly, women were more likely to report high or very high levels of application (72.2 percent) compared to men (58.7 percent).

While the level of application is a subjective measure, changes in reported levels were compared with responses on whether fellows had applied learning to new or existing research over the past 12 months. When fellows reported changes in their level of application, but they did not report any specific action, further clarification was sought. This is illustrated by the quote below in which a participant describes how her desire and personal initiative to apply what she learned during the course were limited by unfavorable research conditions.

> *[M]y research prior to the GREAT course was not sufficiently gendered and the GREAT course allowed me to anticipate research questions that can be used to fill this gap. I have not been able to deepen these research questions [due to] laboratory constraints and the stage of the project, but I am still thinking about it. I am researching participatory selection criteria, including gender, that will allow me to do more in-depth analysis.*
>
> (Key informant interview, female geneticist breeder, 2020, Burkina Faso)

Triangulating responses in this manner both increases the reliability of quantitative responses and provides contextualization on the different ways in which fellows are applying their learning (Mentz 2017). Descriptions of the most significant changes that fellows made were also coded and analyzed. Coding responses enabled a systematic classification of change stories and made it easier to detect patterns in their narratives. The analysis, presented in Table 7, provided some useful insights on the areas in which fellows felt that they had consistently made the most significant changes. For example, shifts in methods for data

collection with more deliberate targeting and engaging both men and women in research activities were among the most common narratives reported; conversely, references to resource allocation and monitoring and evaluation of agricultural research on gender were very limited.

### *Use of knowledge, skills, and approaches to support institutional actions (outcome monitoring)*

GREAT's approach to institutional engagement has focused on training a "critical mass" that collectively advance GRR through peer-to-peer support, awareness raising, and changes to how research is conducted. Outcome monitoring sought to test this assumption by analyzing what institutional actions, if any, were taken at least one year after the course. Table 8 outlines the most commonly reported actions, broken down by cohorts, gender, and discipline. The data show that, irrespective of their discipline, men were more likely to advocate for changes in research practices or guidelines compared to women, while women were more likely to provide trainings or participate in the drafting of gender policies or strategies compared to men. Sharing resources was the most commonly reported action, with 82 percent of female biophysical scientists reporting having done this.

While a strong interest in sharing information and learning was apparent, outcome monitoring also provided some insights on the most significant barriers to application. Table 9, below, provides an overview of the outcome data collected from participants in the customized GREAT courses. Despite small sample sizes, disaggregation by gender and discipline revealed some interesting preliminary patterns. Convincing others of the value of GRR was one of the most common challenges reported among social scientists (specifically, by 53 percent of social scientists compared to 26 percent of biophysical scientists); this challenge was also more pressing for men than for women. Using GRR insights to inform changes in research activities was the most common challenge

| Institutional actions | 2018 (T2) | 2019 (T3) | 2020 (T4) | Bio (F) | Soc (F) | Bio (M) | Soc (M) |
|---|---|---|---|---|---|---|---|
| Shared resources from GREAT with colleagues or supervisors | 79% | 80% | 75% | 82% | 74% | 64% | 77% |
| Provided training or gender sensitization meetings | 17% | 70% | 38% | 55% | 52% | 30% | 23% |
| Advocated for changes in research practices or guidelines to be more gender-responsive | 63% | 60% | 69% | 45% | 52% | 61% | 62% |
| Participated in the drafting of gender policies or strategies in their institutions | 21% | 10% | 13% | 18% | 30% | 9% | 8% |

**Table 8** Institutional actions reported by fellows, in percentages. *(Source: Outcome monitoring data, 2018-2020.)*

| Top reported challenges | All bio | Bio (F) | Bio (M) | All Soc | Soc (F) | Soc (M) |
|---|---|---|---|---|---|---|
| Convincing others of the value of gender-responsive research (GRR) | **26%** | 17% | 33% | **53%** | 44% | 67% |
| Understanding how to integrate GRR into research design | **33%** | 33% | 33% | **20%** | 22% | 17% |
| Developing GRR data collection tools and strategies | **37%** | 25% | 47% | **53%** | 67% | 33% |
| Conducting mixed methods data analysis | **44%** | 33% | 53% | **40%** | 33% | 50% |
| Using GRR insights to inform changes in research activities | **52%** | 42% | 60% | **27%** | 44% | 0% |
| Integrating gender in breeding program activities | **19%** | 17% | 20% | **33%** | 11% | 67% |
| Communicating effectively about GRR results | **19%** | 17% | 20% | **33%** | 33% | 33% |
| Sample (N) | **27** | 12 | 15 | **15** | 9 | 6 |

**Table 9** Key challenges reported by course participants, in percentages. *(Source: Outcome monitoring data for participants of customized courses, 2020.)*

among biophysical scientists (52 percent) yet was less prominent among social scientists (27 percent) and particularly male social scientists, of which none in the sample reported this as a top challenge.

Annual data collection illustrated the kinds of changes that fellows were making over time and the barriers to enacting GRR. However, the data did not provide quality assessments of the research practices reported or of other factors that contributed to improving GRR practices and products. This information was sought through the case studies discussed below.

### *Case studies to examine GRR implementation*

In the case studies, quality of research[10] was evaluated against the criteria constituting "good" GRR practices. The criteria were informed by GREAT's fieldwork checklist for appraising research quality on gender integration into research design, data collection, analysis, and presentation. For each of the above criteria, a detailed guide of how the research would be appraised and examples of good and bad practice were developed by a gender consultant to enable a consistent application of the assessment. Fellows were also asked about their specific role in the production of different research products, whether other gender specialists or organizations had been involved and, if so, at what stage. The criteria are presented in Table 10.

Figure 7 shows the number of teams that met the criteria of good GRR practices in research design, collection, analysis, and presentation. This provides an independent account of the degree of progress in terms of actual application. For example, in the design stage, research more explicitly outlined the methods for how GRR would be conducted and what gender-responsive research questions would be answered. However, in most cases the link to how the research would contribute to gender equality objectives and outcomes was not clearly articulated. In terms of data analysis and presentation, half of the case studies were unable to produce much more than descriptive analysis. While revisions to research design, collection, analysis, and presentation demonstrate initiative and a conscious effort to capture differential experiences, perspectives, and needs in a structured manner, there were nonetheless clear limitations in how research was being designed and implemented to contribute to more transformative outcomes. Alongside the appraisal of research quality, case studies provided important context on what skills were particularly valuable several years after completing the course and on how these skills varied depending on researchers' different roles and responsibilities. For example, several researchers in supervisory roles spoke in more general terms about increased awareness of gender issues and exposure to new participatory methods. Understanding the complexities of GRR and what is required to undertake it enabled them to create the conditions for their teams to implement the techniques more effectively.

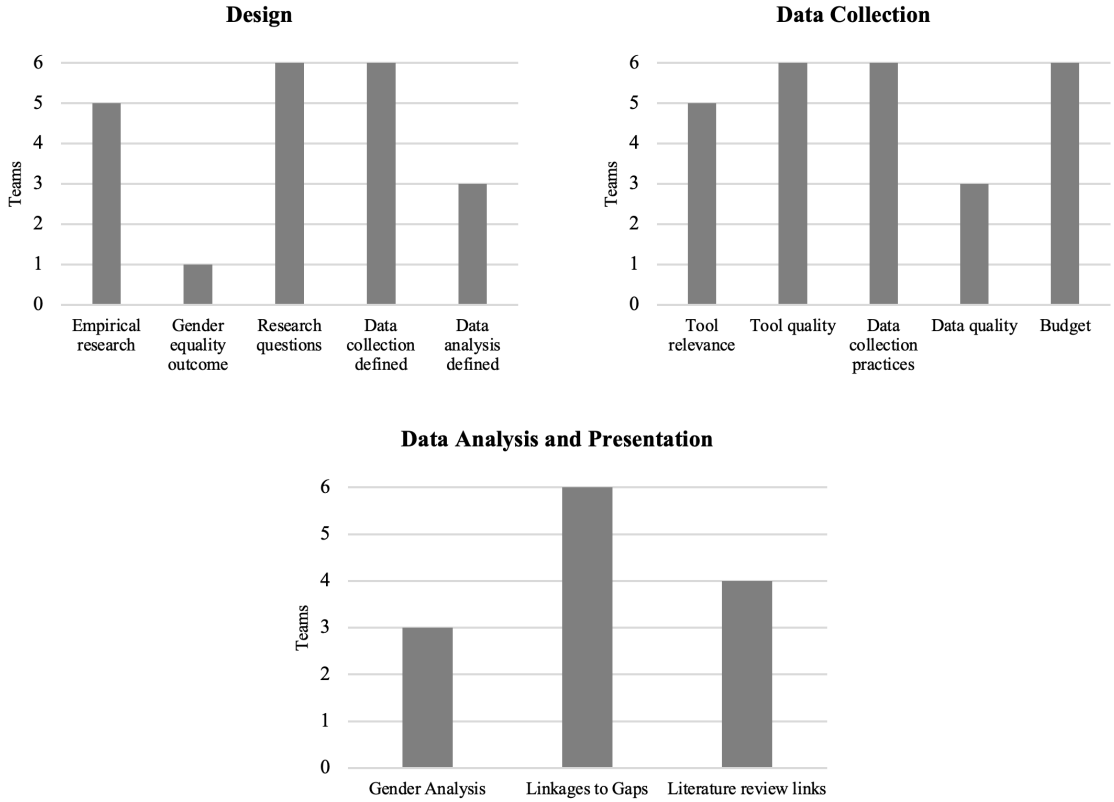


**Figure 7** Appraisal of fellows' research against set criteria.

| **Research design** | |
|---|---|
| **Citation of previous research** | Citation of previous research on gender issues to provide background that explains the need for the research in the proposal. The background is expected to present descriptions from scholarly literature or previous studies or data sets at the institutions of men's and women's division of labor in agriculture, preferred varieties, participation in seed selection, involvement in marketing/trade, and/or access to key productive resources. |
| **Gender equality outcome(s)** | Presence of a clear statement of the expected gender equality outcome based on the description of the critical and relevant gender disparities that were considered in the project. |
| **Gender-related research questions** | Presence of a clear statement of the gender-related research question(s). |
| **Methodology (data collection)** | The methodology specifies clearly what types of gender-related data will be collected and how. It clearly specifies that data will be collected from women and men, and it also pays attention to other social variables such as age, ethnicity, and income. |
| **Methodology (data analysis)** | The methodology/work plan describes how the gender analysis will be done, by whom, and whether it will be qualitative or quantitative. It also identifies the specific software to be used and the types of questions that might be used to interrogate the data. |
| **Data collection** | |
| **Question design** | The topics covered in the data collection tools clearly connect to the gender-related research question(s). |
| **Tool design** | Tools (surveys, interview schedules, etc.) reflect good practices of qualitative and quantitative question design and accurately capture both men's and women's information. |
| **Data collection practice** | The data collection processes adhere to good practices of gender-responsive data collection (e.g., interviewing men and women separately, sample sizes that ensure representation of men and women). |
| **Sex-disaggregated datasets** | Fellows collect and report on sex-disaggregated data for men and women through the use of tailored quantitative and qualitative tools. |
| **Budget** | There is sufficient budget included to carry out gender-responsive data collection and analysis. |
| **Data analysis and presentation** | |
| **Gender analysis** | The research includes a gender analysis in which gender dimensions are highlighted in the analysis. |
| **Link to problem analysis** | The analysis shows how the findings address the information gap identified during the conceptualization. |
| **Appropriate context** | The analysis is backed up by a solid literature review on gender issues linked to both the research question(s) and the local context in which the research is conducted. |
| **Use of research results** | |
| **Communication** | Results are presented in different communication forms and shared with colleagues as well as farmers and participants of the research. |
| **Use case** | Results are used to inform practice changes in the breeding program. |

**Table 10** Criteria used to assess good practices of gender-responsive research.

This contrasts with more technical references to gender-related skills in data analysis, presentation, and report writing from several social science researchers who regularly contribute technical skills to gender-related data collection and analysis within their research teams. Across all case studies, social scientists reported greater confidence and increased opportunities to carry out gender work. Some also went so far as to state their services have been in greater demand likely as a consequence of the parallel training of biophysical researchers. References to increased collaboration in the development of research proposals and implementation from both participants and organizational leadership also provided further insights on the ways in which the team approach had catalyzed more interdisciplinarity in other research. This helps the training team to better understand how participants in varied roles may benefit differently and it confirms the value of training teams, despite the greater time commitments and expenses for institutions to send multiple researchers.

Information on the institutional context was also sought through the case studies. For example, gender policies and/or guidelines were reviewed. While there was some form of gender policy for all focus institutions, in the majority of cases researchers noted policies were not specific enough to usefully inform their work. Nonetheless, the existence of such policies and guidelines, be it at the organizational level or required by a funder, created an important incentive for gender to be considered. However, more comprehensive GRR (i.e., including practices that go beyond the collection of gender-disaggregated data) was only carried out where there was institutional support and dedicated resources to do so. Higher performing teams came from institutions which not only had larger projects, placing greater attention to gender already at the start of the course, but also dedicated resources for further implementation of GRR. This highlights that if an organization is not facilitating changes in the habits and practices of its researchers, organizational performance is unlikely to substantially improve, even if individuals within that organization have acquired the right skills and capacities. This is a particularly important point of consideration for the training team, both in terms of the recruitment strategy as well as the development of post-training support and alignment with other capacity development initiatives.

Lastly, case studies provided evidence of how results were used and of whether there were any indications of positive outcomes for project beneficiaries. Fellow responses were triangulated with feedback from other research colleagues, supervisors, and beneficiaries of the research, such as farmers involved in on-farm trials. The excerpts below highlight the kinds of information provided on how practices were changing and the preliminary outcomes for those involved.

> *The focus on gender integration during the on-farm trials is a recent shift in approach. We consider sex of the host farmers for the trials. We deliberately make sure [that] we select fields belonging to women and men, so as to involve both in evaluating the varieties.*
>
> (Interview, male agronomist and colleague who works with GREAT fellows on the research project, 2019, Ghana)

> *Now we are involved in the on-farm trials. Previously researchers would bring the varieties and give [them] to individual farmers who were in most cases men to plant. In such cases, the field would always be managed by the man, who would then do the evaluation individually, with no opportunity to hear views of other farmers, especially us women. The new process has enabled us to have a voice in evaluating and selecting the variety of choice thus accept[ing] the new variety and adopt[ing] it to increase production.*
>
> (Focus group discussion, woman, 2019, Ghana)

Information from the case studies provided necessary context to the self-reported information generated from the annual outcome surveys. They gave greater insights on how different skills were valued and applied to diverse research environments and what kinds of benefits, if any, were materializing. Using the GREAT project team as the unit of analysis also provided useful insights on how interdisciplinary collaboration continued from assignments and lessons initiated during the course.

## Discussion

The GREAT program invested significantly in monitoring, learning, and evaluation of the course consistently over time, making it a critical component of its program's implementation. The MLE approach sought to build learning and participatory approaches within different methods and time-points in order to capture the complexity of capacity development, while balancing different perspectives and triangulating where possible. As underscored by the literature, assessments of training contributions to capacity development require longer time horizons, flexibility, and exploration of changes at different levels. While the program made positive progress towards evaluating changes in fellows' capacity to conduct GRR, there were nonetheless limitations to the approach. This section presents some reflections on the strengths and challenges of implementing the MLE system.

### *Learning-centric and participatory approaches to improve program design*

Continual opportunities for learning- and data-driven adaptations in the course programming were central to measuring the capacity development of participants. GREAT is an experimental course seeking to test out teaching methods and approaches for GRR. In response, a utilization focused approach that prioritized learning was very important.

This happened at several points over the course delivery and during the five-year period. At the design stage, the development and use of evaluation and learning questions on a core set of issues enabled regular reflection on what the course was aiming to achieve, what evidence was available, and what was still needed. Regular discussions were needed to clearly define terms of measurement, such as capacity itself, as well as to adjust the structure of the program. Interdisciplinary teams and diverse experiences, roles, and knowledge of gender required responsive programming dependent on measurements at the personal level

and relational level within research teams. For example, it was important to understand individual technical competencies for conducting GRR as well as capture the team settings and the quality of research that was produced through interdisciplinary collaboration.

The learning loops and annual workshops (one to two days long) played a key role in fostering more strategic reflection on the trajectory of the program, what needed to be changed and on what basis. The collaborative review by the MLE and program team improved session content and delivery methods through iterative feedback. Further, curriculum review meetings allowed trainers to reflect on their own delivery and to provide direct feedback to each other, building and reflecting on both trainers' and participants' capacities (see Mangheni et al., this issue). Suggestions (if any) for further improvements were provided to the trainers for another iteration on session content before a final review by the program management team.

Participants' voices were included in the regular data collection process and directly contributed to refinements in program design and delivery. However, while participants in the pilot course were involved in developing the program's theory of change, this process was not repeated in subsequent courses due to time limitations. Regularly engaging participants of GREAT in the creation and refinement of the initiative's theory of change as part of the training could help to enhance fellows' understanding of their own roles, refine the assumption made about how change happens post-course, and engender a greater sense of responsibility for sharing information and institutional progress. It could also enable a more realistic representation of their personal aspirations in attending the course and the resultant outcomes.

Another takeaway is the need to balance the time and effort in measuring learning and experiences during the course with measurement efforts afterward. The experimental nature of the course meant that understanding its quality was important, but tools to capture that needed to be light touch and practical. For example, we moved away from structured daily feedback to capturing what was working and what wasn't on shorthand sticky notes. Additionally, adult learners approach education differently than traditional students and as observed by Meyer (2002) tend to expect immediate feedback and critical evaluations that can help them navigate real-world, complex situations. In this sense "assessment for learning" rather than "assessment of learning" becomes particularly important and can render the more typical pre-post assessments less relevant (Bin Mubayrik 2020). A careful balance is needed that captures experiences and growth and provides useful feedback while not being overly cumbersome.

### *Multi-stage data collection using different methods*

The MLE system was built on the idea of capturing learning and practice change at different time points. Opportunities for researchers to apply concepts and techniques learned through GREAT often depended on the status of the research projects within their home institutions. As such, it was critical to allow sufficient time to pass while still aiming to understand what kind of momentum was generated shortly after the course and conversely whether technical limitations noted at the end of the course hindered application. Furthermore, following up with researchers after some time had passed meant they had been exposed to different scenarios without the technical assistance of course trainers. They could then speak about their own limitations and/or identify other technical assistance or support needed with greater clarity. This was complemented by in-depth case studies which illuminated what factors helped or hindered translation of training into practice change within teams and breeding programs. For example, findings helped to confirm the importance of linking training to on-going projects and the value of recruiting research teams with participants in decision-making roles. The benefits derived from training multiple teams of fellows from the same institution ultimately informed targeting and recruitment strategies.

However, reliance on self-reported data from the outcome monitoring surveys was a primary limitation of the approach. While self-reported assessments continue to be one of the most used tools for collecting large amounts of data (Borden and Zak-Owens 2001) there are limitations of these measures. Although efforts were made to triangulate findings with verifiable data and accounts by supervisors, in practice it was very difficult to access researchers' referenced data or proposals due to concerns over sharing proprietary information. This challenged objective assessments of quality and limited the depth of fellows' post-training GRR work. It also meant we had limited insights as to how practices were being integrated across the entire research cycle. For example, there were very few references from fellows on increased resource allocation for GRR or specific examples of how gender-related indicators were being tracked. Whilst this points to areas where fellows were less able to make changes, without access to budgets or evaluation frameworks, these areas cannot really be assessed. Standard metrics and frameworks that institutions themselves can apply (see Mangheni et al., this issue) could help to create more standardized measures of GRR that would enable funders, researchers, and other development practitioners to track progress. This, however, requires building greater institutional engagement, making the case for its value, and setting out stronger expectations for information sharing at the outset of training engagement.

On the other hand, the case studies provided very rich and detailed insights on fellows' contexts. They help to illuminate not just whether there had been short-term improvements in individual or organizational performance, but how and where teams were able to make positive inroads, and how skills were valued and deployed differently depending on their unique roles, starting points, and access to resources. For example, in one of the case studies where more modest improvements were reported, understanding the relative isolation from broader African scientific networks, major funding constraints, and lack of PhD-level staff available in the country were important factors to acknowledge if more progress is to be made and sustained. In this sense, it was critical to ensure that "success" was not narrowly viewed against a standard metric but also considered in light of researcher and institutional starting points. However, the small sample size and qualitative nature of the case studies meant that responses were

illustrative, but not generalizable to the wider group of training participants. This also meant that in-depth insights were not captured post-course for those participating in later versions of the course. As the GREAT course underwent significant changes after its second year, additional case studies should be planned and budgeted for to understand the potential implications of changes in the model to the application of learning. Further analysis of institution-wide changes and the extent to which social outcomes materialize from changing practices over longer time horizons will also help to make a stronger case for further investments in GRR.

## Future considerations

The experience of implementing the GREAT MLE system provides important lessons for practitioners considering approaches to measure and assess complex, interdisciplinary gender-responsive training programs with participants from very diverse research contexts. These are summarized below.

*Ensure that adequate financial and human resources are dedicated to developing and implementing the MLE system.* Significant time and resources are required to successfully implement a thoughtful and rigorous MLE system and should be planned for and budgeted upfront. In particular, developing appropriate data collection tools, systems, and learning processes necessitates a certain level of technical training by staff responsible for MLE implementation and a significant investment of time for the staff more broadly. Program designs should consider both the financial resources required for implementation as well as who will be responsible for implementation (i.e., whether the system will be mainly delivered internally or partly delivered through external support). Such decisions should be thought through and the trade-offs of internal and external support weighed carefully.

*Engender an evidence- and learning-based approach among partners and participants.* This includes clearly defining at the outset what the program hopes to learn, how the information will be captured, and what processes are necessary to ensure that the information can be used to inform adaptive management. One method is to engage capacity development participants in the creation and refinement of the initiative's theory of change. Similarly, tracking whether capacity development initiatives are contributing to system-wide changes more broadly requires coordinated efforts across research institutions. This is necessary to understand the collective contributions of other drivers intended to catalyze more GRR.

*Acknowledge that change does not happen linearly and can take time depending on the capacity of the team, the stage of research, and the enabling environment of research institutions.* Where possible, shorter term progress markers should be integrated into the monitoring system in order to understand whether the approach is helping to advance changes in the intended direction. This also requires acknowledging that individual capacity is part of a larger ecosystem and is not simply about individual skill attainment.

*Embrace complexity and regularly reflect on the theory of change.* In an experimental course, the way in which capacity development is being delivered is constantly changing. As a result, the assumptions that were made at the outset about the relevant pathways may no longer be valid. It is important to reflect on how shifts in the program and evidence more deliberately influence the overall strategy and the subsequent implications for data collection and learning.

*Document approaches and regularly reflect on the credibility and value of the information generated through those approaches.* It is important to regularly document what is done and how it enhances consistency of its application over time, particularly if approaches are carried out by different staff or evaluators. This also ensures that the data collected are not only useful and credible, but that information is also actively used.

## Conclusion

Developing a monitoring, learning, and evaluation approach for GRR training programs can support the ultimate goal of increasing capacity to integrate and address gender equity through agricultural research and programming. Systematic approaches to capturing and measuring the potential impacts of these programs will not only serve to improve their content, by supporting the development of minimum standards, but they will also help to make a case for their effectiveness and importance. As capacity development is non-linear, occurring at multiple time points and levels, MLE approaches need to be learning-centered and participatory, and they should rely on multiple methods at different time points. While there are some limitations and future considerations, GREAT offers a crucial example of how GRR and related capacity building efforts can be strengthened.

## References

Alkire, S., Meinzen-Dick, R., Peterman, A., Quisumbing, A., Seymour, G. and Vaz, A. (2013) 'The Women's Empowerment in Agriculture Index', *World Development*, 52(C), pp 71–91.

Asian Development Bank (2008) *Effectiveness of ADB's capacity development assistance: how to get institutions right*, Special Evaluation Study SES:REG 2008-05. Mandaluyong, PH: Asian Development Bank (ADB).

Balbach, E. (1999) *Using case studies to do program evaluation*, Guide. Sacramento, CA: California Department of Health Services. Available at https://www.betterevaluation.org/sites/default/files/ProgramEvaluation.pdf

Bamberger, M., Rao, V., and Woolcock, M. (2010) *Using mixed methods in monitoring and evaluation: experiences from international development*, Policy Research Working Paper 5245. Washington, DC: World Bank.

BetterEvaluation (2016) *Randomized Controlled Trial*, Article. Available at https://www.betterevaluation.org/en/plan/approach/rct

Bin Mubayrik, H.F. (2020) 'New trends in formative-summative evaluations for adult education', *SAGE Open*, 10(3), pp.1–13.

Blume, B., Ford, J., Baldwin, T., and Huang, J. (2010) 'Transfer of training: a meta-analytic review', *Journal of Management*, 36(4), pp. 1065–1105.

Borden, V. and Zak-Owens, J.L. (2001) *Measuring quality: choosing among surveys and other assessment of college quality*, 1st edn. Washington, DC: American Council on Education.

Bryan, E., Bernier, Q., Espinal, M., and Ringler, C. (2016) *Integrating gender into climate change adaptation programs: a research and capacity needs assessment for sub-Saharan Africa*, CCAFS Working Paper No. 163. Copenhagen, DK: CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS).

Bustelo, M., Ferguson, L., and Forest, M. (2016) 'Introduction', in Bustelo, M., Ferguson, L., and Forest, M. (eds.) *The politics of feminist knowledge transfer: gender training and gender expertise*. London, UK: Palgrave Macmillan, pp. 1–24.

Cole SM., Kantor P., Sarapura S., and Rajaratnam S. (2014) *Gender-transformative approaches to address inequalities in food, nutrition and economic outcomes in aquatic agricultural systems*, Working Paper AAS-2014-42. Penang, MY: CGIAR Research Program on Aquatic Agricultural Systems (AAS).

Davies, P. (2012) 'Impact evaluations and higher education interventions for development', *Measuring the Impact of Higher Education Interventions for Development*. London International Development Centre, 19–20 March. Available at https://lidc.ac.uk/$_{a}$*ssets/P*%20*Davies.pdf*

EIGE (European Institute for Gender Equality) (2016) *Gender equality training: gender mainstreaming toolkit*. Luxembourg, LU: EIGE.

Estrella, M. and Gaventa, J. (1997) *Who counts reality? participatory monitoring and evaluation: a literature review*, IDS Working Paper 70. Brighton, UK: Institute of Development Studies.

FAO (Food and Agriculture Organization of the United Nations) (2011) *Women in agriculture: closing the gender gap for development*, The State of Food and Agriculture 2010-2011, Report. Rome, IT: FAO.

FAO (Food and Agriculture Organization of the United Nations) (n.d.) *Monitoring capacity development*, FAO Capacity Development Learning Module. Available at https://www.fao.org/capacity-development/resources/practical-tools/monitor-capacity-development/en/

FSG (2015) *Guide to actor mapping*, FSG Tools. Available at: https://www.fsg.org/tools-and-resources/guide-actor-mapping

Guijt, I. (2008) *Critical readings on assessing and learning for social change: a review*, IDS Development Bibliography 21. Brighton, UK: Institute of Development Studies (IDS).

Gutierrez-Montes, I., Arguedas, M., Ramirez-Aguero, F., Mercado, L., and Sellare, J. (2020) 'Contributing to the construction of a framework for improved gender integration into climate-smart agriculture projects monitoring and evaluation: MAP-Norway experience', *Climatic Change*, 158, pp. 93–106.

Hillenbrand, E., Karim, N., Mohanraj, P., and Wu, D. (2015) *Measuring gender-transformative change: a review of literature and promising practices*, Working Paper. Atlanta, GA: CARE USA.

Horton, D., Mackay, R., Andersen, A., and Dupleich, L. (2000) *Evaluating capacity development in planning, monitoring, and evaluation: A case from agricultural research*. Report no. 17. The Hague, Netherlands: International Service for National Agricultural Research.

Howland, F., Acosta, M., Muriel, J., and Le Coq, J.F. (2021) 'Examining the barriers to gender integration in agriculture, climate change, food security, and nutrition policies: Guatemalan and Honduran perspectives', *Frontiers in Sustainable Food Systems*, 5, Article 664253.

INTRAC (International NGO Training and Research Centre) (2016) *Tracking capacity change*, Report. Available at https://www.intrac.org/wpcms/wp-content/uploads/2016/10/Tracking-Capacity-Change.pdf

INTRAC (International NGO Training and Research Centre) (2017a) *Theory-based evaluation*, M&E Training and Consultancy Resource. Available at https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Theory-based-evaluation.pdf

INTRAC (International NGO Training and Research Centre) (2017b) *Utilisation-focused evaluation*, M&E Training and Consultancy Resource. Available at https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Utilisation-focused-evaluation.pdf

Klatt, J., and Taylor-Powell, E. (2005) *Using the retrospective post-then-pre design*, Program Development and Evaluation Quick Tips 27. Madison, WI: University of Wisconsin–Extension.

Kristjanson, P., Bryan, E., Bernier, Q., Twyman, J., Meinzen-Dick, R., Kieran, C., Ringler, C., Jost, C., and Doss, C. (2017). 'Addressing gender in agricultural research for development in the face of a changing climate: where are we and where should we be going?', *International Journal Agricultural Sustainability*, 15(5), pp. 482–500.

Lombardini, S., Bowman, K., and Garwood, R. (2017) *A "how to" guide to measuring women's empowerment: sharing experiences from Oxfam's impact evaluations*, Report. Oxford, UK: Oxfam.

Lam, T.C. and Bengo, P. (2003) 'A comparison of three retrospective self-reporting methods of measuring change in instruction practice', *American Journal of Evaluation*, 24(1), pp. 65–80.

Mangheni, M.N., Tufan, H.A., Boonabana, B., Musiimenta, P., Miiro, R., and Njuki, J. (2019) 'Building gender research capacity for non-specialists: lessons and best practices from gender short courses for agricultural researchers in sub-Saharan Africa', in Marcia, T.S., Kristy, K. and Vasilikie, D. (eds.) *Gender and practice: knowledge, policy, organizations*. Bingley, UK: Emerald Publishing, pp. 99–118.

Mangheni, M.N., Musiimenta, P., Boonabaana, B., and Tufan, H.A. (2021a) 'Tracking the gender responsiveness of agricultural research across the research cycle: a monitoring and evaluation framework tested in Uganda and Rwanda', *Journal of Gender, Agriculture and Food Security*, 6(2), pp. 58–72.

Meinzen-Dick, R., Quisumbing, A., Berhman, J., Biermayr-Jenzano, P., Wilde, V., Noordeloos, M., Ragasa, C., and Beintama, N. (2011) *Engendering agricultural research, development, and extension*. Washington, DC: International Food Policy Research Institute (IFPRI).

Mendizabal, E., Datta, A., and Young, J. (2011) *Developing capacities for better research uptake: the experience of ODI's Research and Policy in Development Programme*, Background Note. London, UK: Overseas Development Institute (ODI).

Mentz, M. (2017) 'The benefits of both worlds: towards an integrated mixed-methods approach for evaluating women's empowerment', *Journal of Gender, Agriculture and Food Security*, 2(1), pp.14–34.

Meyer, K. (2002) *Quality in distance education: focus on online learning*, ASHE-ERIC Higher Education Report, Volume 29(4). San Francisco, CA: Jossey-Bass.

Miles, M.B., and Huberman, A.M. (1994) *Qualitative data analysis: an expanded sourcebook*, 2nd ed. Thousand Oaks, CA: SAGE.

Morgan, P. (2006) *The concept of capacity*, Capacity, Change and Performance Report. Maastricht, NL: European Centre for Development Policy Management (ECDPM).

Mukhopadhyay, M. (2014) 'Mainstreaming gender or reconstituting the mainstream? Gender knowledge in development', *Journal of International Development*, 26(3), pp. 356–367.

Nelson, M. (2006) *Does training work? Re-examining donor-sponsored training programs in developing countries*, Capacity Development Brief No. 15. Washington, DC: World Bank.

Njuki, J. (2016) 'Practical notes: critical elements for integrating gender in agricultural research and development projects and programs', *Journal of Gender, Agriculture and Food Security*, 1(3), pp. 104–108.

Njuki, J., Eissler, E., Malapit, H., Meinzen-Dick, R., Bryan, E., and Quisumbing, A. (2021) *A review of evidence on gender equality, women's empowerment, and food systems*, United Nations Food Systems Summit 2021 Brief. Washington, DC: International Food Policy Research Institute (IFPRI).

Nowell, L., Norris, J., and White, D. (2017) 'Thematic analysis: striving to meet the trustworthiness criteria', *International Journal of Qualitative Methods*, 16(1), pp.1–13.

ODI (Overseas Development Institute) (2009) *Strategy development: Most Significant Change (MSC)*, ODI Tools for Knowledge and Learning. Available at https://odi.org/en/publications/strategy-development-most-significant-change-msc/

Ortiz, A. and Taylor, P. (2009) *Learning purposefully in capacity development*, IIEP Opinion Paper, Rethinking Capacity Development Series. Paris, FR: International Institute for Educational Planning (IIEP).

Otoo, S., Agapitova, N., and Behrens, J. (2009) *The capacity development results framework: a strategic and results-oriented approach to learning for capacity development*, Working Paper. Washington, DC: World Bank.

Patton, M. (2008) *Utilization-focused evaluation*, 4th edn. Thousand Oaks, CA: SAGE.

Pearson, J. (2011) *Training and beyond: seeking better practices for capacity development*, OECD Development Co-operation Working Paper No. 1. Paris, FR: OECD.

Preskill, H. and Boyle, S. (2008) 'A multidisciplinary model of evaluation capacity building', *American Journal of Evaluation*, 29(4), pp. 443–459.

Rubin, D. (2016) *Qualitative methods for gender research in agricultural development*, IFPRI Discussion Paper 01535. Washington, DC: International Food Policy Research Institute (IFPRI).

Sarapura Escobar, S. and Puskur, R. (2014) *Gender capacity development and organizational culture change in the CGIAR Research Program on Aquatic Agricultural Systems: a conceptual framework*, Working Paper AAS-2014-45. Penang, MY: CGIAR Research Program on Aquatic Agricultural Systems (AAS).

SDC (Swiss Agency for Development and Cooperation) (2012) 'Monitoring and evaluating empowerment processes', in OECD (ed.) *Poverty reduction and pro-poor growth: the role of empowerment*. Paris, FR: OECD.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., and Befani, B. (2012) *Broadening the range of designs and methods for impact evaluations*, DFID Working Paper 38. London, UK: Department for International Development (DFID).

Taylor, P. and Clarke, P. (2008) *Capacity for a change: from the capacity collective workshop*, IDS Report. Brighton, UK: Institute of Development Studies (IDS).

ten Cate, T.J. and De Haes, J.C.J.M. (2000) 'Summative assessment of medical students in the affective domain', *Medical Teacher*, 22(1), pp. 40–43.

Tufan, H.A., Mangheni, M.N., Boonabaana, B., Asiimwe, E., Jenkins, D., and Garner, E. (2021) 'GREAT Expectations: building a model for applied gender training for crop improvement', *Journal of Gender, Agriculture and Food Security*, 6(2), pp. 1–18.

Vallejo, B. and Wehn, U. (2016) 'Capacity development evaluation: the challenge of the results agenda and measuring return on investment in the Global South', *World Development*, 79, pp. 1–13.

Vogel, I. (2012) *Review of the use of "Theory of Change" in international development*, Review Report. London, UK: Department for International Development (DFID).

Walby, S. (2005) 'Gender mainstreaming: productive tensions in theory and practice', *Social Politics: International Studies in Gender, State and Society*, 12(3), pp. 321–343.

Wilson, K. (2015) 'Towards a radical re-appropriation: gender, development and neoliberal feminism', *Development and Change*, 46 (4), pp. 803–832.

Woolcock, M. (2009) 'Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy', *Journal of Development Effectiveness*, 1(1), pp. 1–14.

## Notes

1. We use the term monitoring, learning and evaluation (MLE), rather than the more commonly used acronym MEL (monitoring, evaluation and learning) to underscore the importance of the learning process. We believe that learning is part of an active and participatory observation and reflection. MEL is only used when cited in the literature.
2. This article uses the term capacity development rather than capacity building to place greater emphasis on the agency of those being trained, acknowledging that knowledge and skills rest within training participants and that capacity development is a two-way street.
3. We use here the GREAT's definition of gender-responsive research, according to which gender-responsive research involves the use of social science theories, concepts, methods, and tools to investigate the different needs, priorities, and constraints of both men and women so as to address and reduce them, rather than exacerbating any existing gender inequalities (Rubin 2016).
4. In 2021, GREAT delivered the "Theme 5, Gender Responsive Plant Breeding" virtual training course. Data from this course have not been included in this article since outcome monitoring and analysis had still not taken place at the time of writing.
5. One team had to be excluded because team members were unresponsive to repeated requests for interviews and documents over several months.
6. One team had to be excluded because team members were unresponsive to repeated requests for interviews and documents over several months.
7. Levels were self-classified based on management authority.
8. Sex- and discipline-disaggregated data did not show statistically significant differences.
9. Mentors were assigned to each team and they provided technical guidance during the field research.
10. Case studies reviewed their research as part of the GREAT course. This research was carried out both during and after the completion of the course.